Zdalny dostęp do Statystycznych Baz Danych a bezpieczeństwo danych jednostkowych.

Przegląd zastosowanych rozwiązań urzędów statystycznych na świecie.

mgr inż. Jarosław Butanowicz
mgr inż. Łukasz Ślęzak

Klasyfikacja Baz Danych

Bazy danych możemy podzielić na:

- **Faktograficzne bazy danych** - są to uniwersalne bazy danych szeroko stosowane do składowania oraz udostępniania danych.

- **Statystyczne bazy danych** są wykorzystywane do przechowywania danych, które są udostępniane użytkownikom wyłącznie na poziomie statystyk, a nie danych jednostkowych.

Problem ochrony baz danych

W przypadku baz danych ochronę bezpieczeństwa informacji rozumiemy jako kombinację usług zapewniających poufność, integralność i dostępność danych.

- **Poufność** jest właściwością danych, wskazującą obszar, w którym te dane nie powinny być dostępne lub ujawniane nieuprawnionym osobom, procesom lub innym podmiotom.
- **Integralność** jest właściwością polegającą na tym, że dane nie zostały wcześniej zmienione lub zniszczone w nieautoryzowany sposób.
- **Dostępność** jest właściwością danych polegająca na tym, że mogą one być dostępne i wykorzystywane na żądanie uprawnionej jednostki.

Problem ochrony statystycznych baz danych

W systemach zarządzania bazami danych ochrona obejmuje trzy rodzaje sterowania:

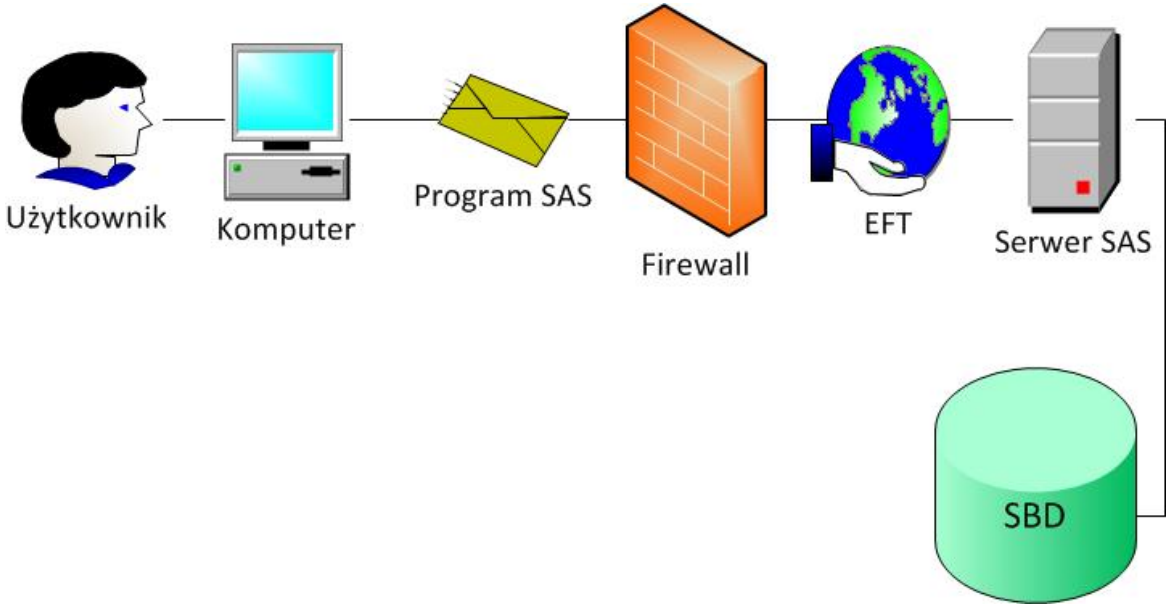
- Sterowanie dostępem,
- Sterowanie przepływem informacji,
- Sterowanie wnioskowaniem.

Zdalny dostęp do SBD

- Z punktu widzenia wygody użytkownika najwygodniejszym sposobem dostępu do SBD jest możliwość uruchomienia analizy z własnego komputera przy uzyskaniu wyników w czasie rzeczywistym charakteryzujących się niewielką utratą zdolności informacyjnej i wysoką precyzją. Tradycyjne sposoby dostępu do danych stosowane w większości Narodowych Urzędów Statystycznych (predefiniowane pliki z ograniczoną porcją informacji, bezpieczne centrum i zdalne przetwarzanie danych) idą na kompromis w co najmniej jednym z tych aspektów. Rozwiązaniem spełniającym zaprezentowane wymagania jest zdalny dostęp do danych.

Statistics Canada – aplikacja Real Time Remote Access (RTRA)

RTRA to aplikacja służąca do zdalnego dostępu do serwera z danymi statystycznymi poprzez internet. Użytkownik ma możliwość korzystania z usługi 24 godziny na dobe, 7 dni w tygodniu.

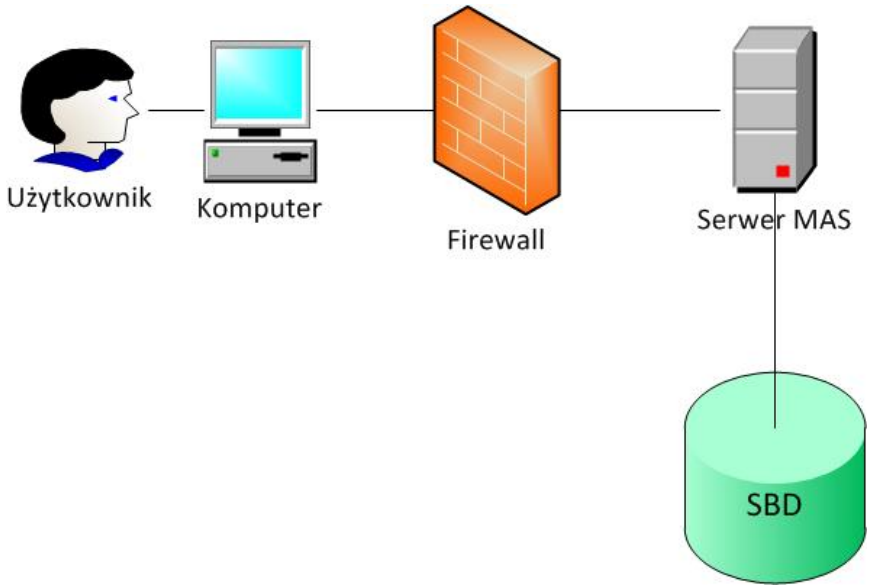


RTRA – bezpieczeństwo

- Ograniczona liczba zapytań. Maksymalnie 10 zapytań na 24 godziny do maksymalnie 10 tabel mogą być wykonane przez użytkownika.
- Zapewnienie poprawności kodu SAS w ramach zapytania. Zapytanie jest weryfikowane pod kątem zgodności z wymaganiami języka SAS, użycia zmiennych i funkcji.
- Korzystanie z ograniczonej ilości dostępnych procesów SAS, aby móc kontrolować wyniki. Prototyp pozwala stosować tylko zmodyfikowaną wersję SAS PROC FREQ.
- Zniekształcanie odpowiedzi metodą ACROUND.
- Możliwość zmiany odpowiedzi na żądanie.
- Wszystkie generowane wyniki przechowywane bezterminowo dla celów kontrolnych.

U.S. Census Bureau – aplikacja Microdata Analysis System (MAS)

Aplikacja MAS posiada graficzny interfejs, który pozwala użytkownikom stworzenie własnej populacji badawczej, a następnie przeprowadzenie analizy regresji, jak i operacji na tabelach krzyżowych. Prototyp alfa został opracowany z wykorzystaneim języka SAS. Wersja beta została wzbogacona o graficzny interfejs w języku Java.

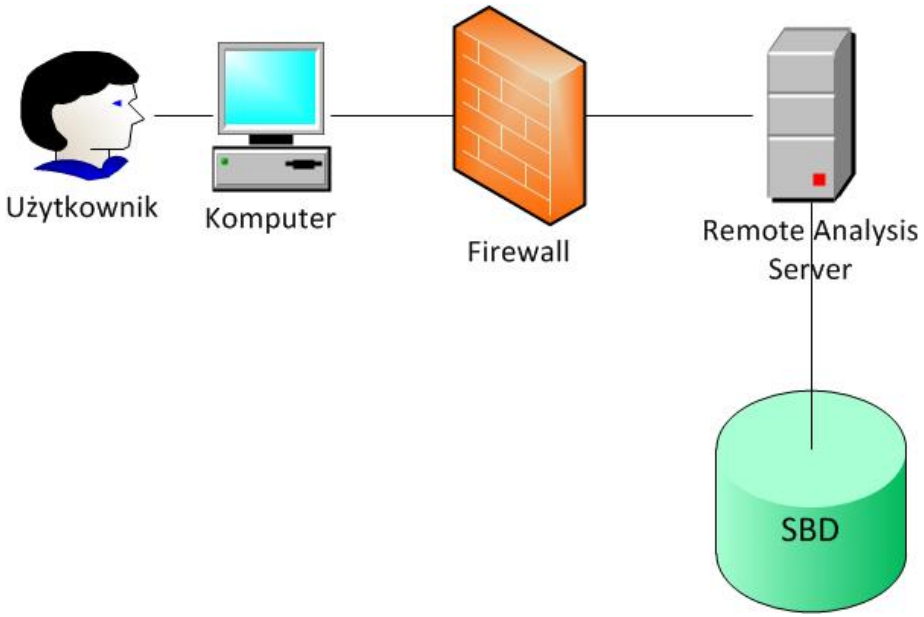


MAS - bezpieczeństwo

- No Marginal 1 or 2 Rule
- Universe Gamma Rule
- Drop Q Rule

Remote Analysis Server – Australian Bureau of Statistics

Aplikacja umożliwiająca zadawanie zapytań do tabel krzyżowych jak i analizy regresji.



RAS - bezpieczeństwo

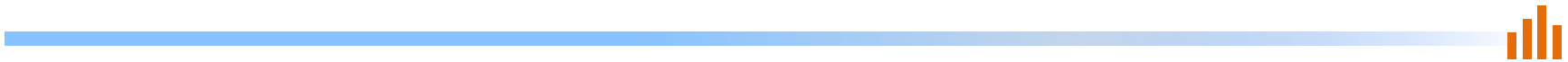
Zniekształcenie danych w tablicach przy użyciu metody spełniającej poniższe kryteria:

- Brak wpływu rozkładu prawdopodobieństwa na wartość komórki tablicy krzyżowej.
- Każda wartość komórki tablicy krzyżowej ma stała wariancję.
- Różniczkowanie dwóch komórek tablicy krzyżowej dla tego samego atrybutu nie usuwa efektu zniekształcenia.
- Przeciwdziałanie usuwaniu zniekształcenia poprzez ataki uśredniające (powtarzanie tego samego zapytania do bazy danych).

Bibliografia

- Šmid W., Metamarketing, Wydawnictwo Profesjonalnej Szkoły Biznesu
- PN-I-02000: Zabezpieczenia w systemach informatycznych – Terminologia, PKN, 1998.
- PN-ISO/IEC 2382-8:2001
- PN-I-02000: Zabezpieczenia w systemach informatycznych – Terminologia, PKN, 1998.
- PN-ISO/IEC 2382-8:2001
- Bleninger, P., Drechsler, J. and Ronning, G. (2010) Remote Data Access and the Risk of Disclosure from Linear Regression: An Empirical Study, Privacy in Statistical Databases,Springer.
- Chipperfield, J. O. and O'Keefe, M. C. (2011), Disclosure-Protected Inference using Generalised Linear Models, Submitted for publication.
- Mathews, G, J. And Harel, O. (2011), Data Confidentiality: A Review of methods for statistical disclosure limitation and methods for assessing privacy, Statistical Surveys, 5, pp. 1-29.
- James Chipperfield and Frank Yu, Australian Bureau of Statistics, Australia 2011 Protecting Confidentiality in a Remote Analysis Server for Tabulation and Analysis of Data
- Jason Lucero, Michael Freiman, Lisa Singh, Jiashen You, Michael DePersio and Laura Zayatz (2011), The Microdata Analysis System at the U.S. Census Bureau
- Michelle Simard, Statistics Canada (2011), Progress with Real Time Remote Access

Q & A



?